Controlling the multiplicity using weighted T-sum statistic

M. Congedo, L. Finos, F. Turkheimer, F. Pesarin

2003.21

Dipartimento di Scienze Statistiche
Università degli Studi
Via C. Battisti 241-243
35121 Padova

Dicembre 2003

Controlling the multiplicity using weighted T-sum statistic

Marco Congedo
Institute for Research in Informatics and Random Systems (IRISA), Rennes, France
marco.congedo@IRISA.fr

Livio Finos
Department of Statistics, The University of Padova, Italy
lfinos@stat.unipd.it

Federico Turkheimer
Neuropathology Department, Imperial College London, UK
federico.turkheimer@imperial.ac.uk

Fortunato Pesarin
Department of Statistics, The University of Padova, Italy
pesarin@stat.unipd.it

Corresponding Author: Marco Congedo,
IRISA, Beaulieu Campus, 35042, Renens, France.
Tel: +02 99847483
E-mail: Marco.Congedo@irisa.fr

Abstract

We introduce a new test procedure for multiple hypothesis testing based on the permutation space of the sum of test-statistics (t-sum). The underlying combining function is shown to be an instance of a family to which it also belongs the well-known combining function based on the maximum of test-statistics (t-max). After discussing the family-wise error rate and the false discovery rate, two common approaches to the control of the type I error in multiple testing, we consider two further error rates, the stochastic family error and the mean square error model fit estimator. By means of a two large set of simulations we show that besides controlling the family-wise error rate in the weak sense, the t-sum procedure also controls the stochastic family error and could considerably outperform the t-max procedure in power and mean square error in experiments with low degrees of freedom. They are also shown several circumstances in which it fits the model better than a procedure controlling the false discovery rate and even better of simply performing a series of univariate tests, which do not control any errors. The t-sum procedure is suitable for pilot and exploratory studies in neuroimaging and in other experimental contexts in which the sample size/ number of hypotheses ratio is low, the data correlation is moderate, and the proportion of false hypothesis is possibly large. We end the discussion outlining possible investigations of the more general form of combining function (weighted sum) with the aim of data-driven selection of an optimal power combining function.

Introduction

With the inception of Positron Emission Tomography (PET) and later development of functional Magnetic Resonance Imaging (fMRI), neuroimaging research for both cognitive and clinical investigations received an astonishing advance. Both PET and fMRI provide three-dimensional images of brain metabolic activity. Electromagnetic tomographies are a recent development, providing functional images from brain electromagnetic data, either electroencephalography (EEG) or magnetoencephalography (MEG).

The analysis of functional experiments results in a volume of suitable statistics, each summarizing the hypothesis of interest at the voxel level. Often the researcher's problem is to identify brain locations were an experimental effect emerges strongly and consistently. In statistical terms the problem is to identify a rejection region for m test-statistics suitably controlling the false positive rate, yet maximizing power. This is a distinctive case of the well known multiple testing problem (Hochberg and Tamhane, 1987; Westfall and Young, 1993), the peculiarity being that in neuroimaging the sample size is very small as compared to the number of hypotheses, which are in the order of thousands, tens of thousands, or even hundreds of thousands.

In the following sections we briefly review the two most popular approaches for the control of type I error, the *family-wise error* (FWE) and the *false discovery rate* (FDR). We discuss two further multiple hypothesis error, the *stochastic family error* (SFE) and the model fit estimator *mean square error* (MSE), stressing their relationship with the FDR. Then, after detailing the current permutation approach to hypothesis testing in functional imaging (t-max) we outline a more generalized approach based on the null

space of weighted sum of test-statistic (t-sum). The ensuing two large sets of simulations compares all errors, model fit and power of t-max, t-sum (equal weights), and simple partial (univariate) permutation tests (t-uni), and power and model fit of t-sum and FDR. Finally, we discuss the merit, limits, possible applications, and future directions of the t-sum procedure. In summary, in this article we expand previous investigations outlining a general framework for combining function and their permutation space. We generalize the step-down algorithm to all possible functions based on the sum of test statistics and we study he properties of the unweighted sum combining function by means of monte carlo simulations. Whereas the motivation for this research springs from the functional imaging data, the t-sum procedure we propose might turn useful in other research fields.

The Multiple Testing Problem

In Neuroimaging as in other experimental situations (e.g., human genome experiment), the *Omnibus Hypothesis* H^0 (also called Intersection Hypothesis or Global null Hypothesis) of "no effect anywhere in the brain" is defined as

$$\boldsymbol{H^0} = \cap_i H^0_i, i:\{1...m\} \qquad (1.0)$$

where H^0_i is the hypothesis related to the individual voxel. Control of the $\boldsymbol{H^0}$ rejection rate is named control of the FWE in the *weak sense*. It is seldom of interest per se, since the researcher wishes to report effects acting on specific locations. Since localization power (selection) of the stochastic effect is sought, the control of the FWE is typically

required in the *strong sense*, i.e., under all configurations of the true and false hypotheses tested. For the remainder of the article let m_0 be the number of true null hypotheses and m_1 the number of false null hypotheses ($m_0 + m_1 = m$). Let also v being the number of falsely rejected null hypotheses.

Ignoring the multiple testing problem and testing each hypothesis individually ("partial" or "univariate" tests) at level α leads to $E(v) = m_0\alpha$. Equivalently, $E(v)/m_0 = \alpha$. Under the omnibus hypothesis $m_0 = m$, thus this rate of false rejection corresponds to $(100 \cdot \alpha)\%$. However the more hypotheses are false the more the error reduces. Ignoring the multiple testing problem is generally considered unacceptable in neuroimaging. We will corroborate this opinion showing the besides not controlling the FWE in the weak sense, which per se makes it inappropriate in a multidimensional context, this leads to very high stochastic error.

Controlling the FWE in the strong Sense

A procedure controlling the FWE in the strong sense at the α level ensures that $P(v>0) \leq \alpha$, i.e., the probability to falsely reject even only one hypothesis is less then or equal to α. For instance, testing each hypothesis at level α/m (Bonferroni) controls the FWE in the strong sense. The popular procedures based on random field theory and randomization-permutation of maximal statistics (for a review see Petersson et *al.*, 1999) also control the FWE in the strong sense, but have usually higher power than the Bonferroni method since they account for the spatial correlation of neuroimaging data.

False Discovery Rate

Benjamini and Hochberg (1995) first proposed to consider the FDR, which is defined as the expected proportion of falsely rejected hypotheses. This approach implies a different notion of the error rate. Indicating by s the number of correctly rejected hypotheses, and by r the number of rejected hypothesis, then $r=(v+s)$ and $FDR=E\{v/r\}$. If $r=0$ the FDR is defined to be 0 as no error of false rejection can be committed. The quantity v/r is the proportion of the hypothesis erroneously rejected as normalized to the total number of hypotheses rejected. This quantity is independent from the rejection rate of the test (Swets, 1988; Swets and Pickett, 1982). This is not the case for the FWE, for which the probability $P(v>0)$ tends to increase as the rejection rate increases. Another difference between the two approaches is that while controlling the FWE the specified error is maintained at each experiment, the FDR controls its own error exactly only on the average of infinite identical experiments (Genovese, Lazar, and Nichols, 2002). Indeed the FDR is not a probability but an expected value.

FDR as optimal minmax estimator of the mean of a sparse gaussian vector

A different interpretation of the FDR controlling procedures has been proposed by Abramovich et al. (2000) and agued for the genomic field by Sabatti (2001).
The neuroimaging problem was stated before as consisting in the simultaneous testing of m generic null hypotheses H^0_i. For simplicity of exposition the problem is now stated

alternatively considering the recovery of an m dimensional mean vector μ from one realization consisting in the m voxel values $x_i \sim N(\mu_i, \sigma_i)$. The problem can be also be re-parametrized as $y_i = x_i / \sigma_i \sim N(\theta_i, \sigma)$. We assume the vector of means θ to be sparse with a relative small proportion of non-zero coordinates. In the context of multivariate estimation theory, the aim is to maximize the information recovery, that is to minimize the squared distance between the estimated and true solution. The model estimator MSE is defined as

$$MSE_{(\hat{\theta}_i, \theta_i)} = \frac{n}{m} \sum_{i=1}^{m} (\hat{\theta}_i - \theta_i)^2 \qquad (2.0)$$

thus, given the sparsity of θ_i, a straightforward approach is possible designing a threshold τ such that the estimate can be calculated as

$$\hat{\theta}_i = \begin{Bmatrix} 0 \, if \; y_i < \tau \\ 1 \, if \; y_i \geq \tau \end{Bmatrix}. \qquad (3.0)$$

After the normalization (n/m), MSE equals 1.0 if the model fit is perfect. The worse the fit, the larger the MSE. The model estimator will obviously depend on the unknown vector θ. Abramovich et al. (2000) showed that an adaptive thresholding procedure constructed to control the false discovery rate (FDR) is asymptotically ($m \to \infty$) optimal for the MSE.

FDR as noise estimator for a procedure controlling the stochastic family error

Let us sort in ascending order the probability values of m identical test-statistics so to obtain

$$\boldsymbol{P}=P_{(1)}> P_{(2)} >\ldots P_{(i)} >\ldots>P_{(m)}\ , i:\{1\ldots m\}. \qquad (4.0)$$

We define the SFE as $P(r>m_1)\leq\alpha$, that is, the probability to reject more hypotheses than are false is less than or equal to alpha. Under the Omnibus hypothesis (1.0) SFE=FWE=FDR, that is, like a procedure controlling the FDR, a procedure controlling the SFE controls the FWE in the weak sense. A procedure controlling the FWE in the strong sense will control the FDR and the SFE. On the contrary a procedure controlling the SFE would control the FWE in the strong sense only in the idealistic case of complete absence of noise, a condition not attainable in practice. We will now argue that when $m_1>0$ the SFE will increase as a function of noise. Suppose $m_1\neq0$ and $m_0\neq0$. In the extreme case of no noise, in the sorted vector $\boldsymbol{P}$ (4.0) there will be complete separation of the H^0 and H^1 hypotheses such that the hypotheses H_i will be true for $i\leq m_0$ and false for $i>m_0$. In such an extreme case a procedure controlling the SFE would also control the FWE in the strong sense and the FDR would be very low. As the noise increases, the overlapping between H^0 and H^1 increases too, and so does the SFE. At the other opposite there will be complete overlapping between H^0 and H^1, that is, the effect is indistinguishable from the noise, and the FDR could be as high as around 0.5. Here noise refers to the stochastic dominance of the distribution under H^0 over the distribution under H^1. The noise tends to zero as the sample size or the effect grows. Therefore, for a procedure controlling the

SFE the *actual* proportion of false rejections, given a fixed α, will decrease as a function of the sample size and of the effect. From this perspective the FDR is not a quantity to be controlled but a measure of the noise. Simultaneous control of the FWE in the weak sense and of the SFE can be considered the minimal requirement for a multiple hypothesis test procedure, in the sense that a procedure incapable of recognizing the omnibus hypothesis *or* rejecting more hypotheses than are false more than $(\alpha.100)\%$ of the times is not worth consideration as a multiple testing procedure.

Permutation test procedures

Brain imaging data-analysis is further complicated by possible non-gaussianity (e.g., electromagnetic current density) and correlation in space and time of physiological measurements. The randomization-permutation approach (Blair et *al.*, 1996; Edgington, 1995; Feinstein, 1993; Fisher, 1935; Lunneborg, 2000; Manly, 1997; Pesarin, 2001; Pitman, 1937a, 1937b, 1938; Troendle, 1995, 1996; Westfall & Young, 1993) offers a flexible and robust solution to this multiple comparison problem. It accounts adaptively for the spatial correlation structure and does not assume data gaussianity. Even more appealing, the test-statistics summarizing the effect sought in the experiment have not to be known by sampling theory, allowing broader and possibly more comprehensive investigations. Because of these properties, the randomization-permutation approach has recently received attention in the neuroimaging literature (Arndt *et al.*, 1996;

Belmonte & Yurgelun-Todd, 2001; Blair & Karnisky, 1994; Bullmore et al., 1996; Holmes et *al.*, 1996; Karniski et *al.*, 1994; Nichols & Holmes, 2001).

At a general level a typical functional imaging study involves m hypotheses, one for each voxel. For simplicity we will assume hereafter that all alternative hypotheses are two-tailed. A similar reasoning applies for one-tailed alternatives and the tests can be adjusted accordingly (e.g., Edgington, 1995; Holmes et al, 1996). At each voxel a test-statistic is derived. The corresponding p-value summarizes the evidence against the null hypothesis. Throughout this paper we will indicate by $\boldsymbol{H}$, $\boldsymbol{T}$, and $\boldsymbol{P}$ the sets of respectively null hypotheses, test-statistics, and p-values. All these sets are comprised of m elements and are indexed by the voxel serial number i:{1...m}. We follow the customary notation of including the index in parenthesis if the set is sorted. All sorting operations in this article refers to sorting in *ascending order,* like in (4.0). In order to control the FWE in the strong sense at the α level a fixed rejection threshold τ is sought such that, for each observed test-statistic T_i, the corresponding hypothesis H_i is rejected if $T_i \geq \tau$. Absolute values of test-statistics and non-negative rejection thresholds are considered because we are considering bi-directional tests.

The t-max permutation test

The (single-step) t-max procedure makes use of the Tippett's combining function (Pesarin, 2001, pag. 148) as applied to test-statistics instead of to the corresponding probability values (Pesarin, 2001, pag. 151). The method defines the rejection threshold τ as the 95th percentile of the null space of the maximal T-statistics. Let B be the number of

data permutations and let the $m \cdot B$ permutation space $\boldsymbol{T}^*$ indicate B sets of m absolute values test-statistics obtained permuting the data under the appropriate exchangeability scheme (Edgington, 1995; Manly, 1997; Pesarin, 2001). Let $\check{T}_j^*$ be the maximal T-statistic obtained in the j^{th} permutation, that is, $\check{T}_j^* = \max_i (T_{ji}^*)$, with $i:\{1...m\}$ and $j:\{1...B\}$. In the terms of Pesarin (2001) this is an element of the permutation space of a direct combination function. The null space is then the set $\check{\boldsymbol{T}}^*$, which is comprised of B elements as obtained in the B realizations of the data permutation. τ is the 95th percentile of $\check{\boldsymbol{T}}^*$. For each observed test-statistic T_i, the corresponding hypothesis H_i is rejected if $T_i \geq \tau$. The procedure has been shown to have strong control over the FWE (Blair and Karnisky, 1994; Holmes et *al.*, 1996). Importantly, it adaptively accounts for the spatial correlation structure of the data and does not assume data gaussianity. The test accounts not only for linear dependence in a multivariate normal framework, but for *any* correlation structures, regardless its form and degree. Depending on the experimental design, even the random sampling assumption may be relaxed. For example, tests on the central tendency are valid whenever subjects have been randomly allocated to groups (Edgington, 1995; Lunneborg, 2000) and if weaker distributional assumptions under the null hypotheses are available (e.g., simmetricity of data distribution instead of normality. See Holmes et *al.*, 1996).

Several step-wise versions of the procedure have been proposed with the aim of increasing its power yet maintaining the same characteristics (Blair and Karnisky, 1994; Holmes et *al.*, 1996; Troendle 1995, 1996; Westfall & Young, 1993). Here we review the common step-down all-at-once procedure (Mainly, 1997) and the algorithm to implement it because the procedure we will propose thereafter is analogous.

The step-down procedure is a sequence of as many as m t-max procedures. As in the canonical step-down procedures the hypotheses are sorted in ascending order respect to their observed statistics. Consider henceforth the set T of statistics sorted by magnitude and the rank index $i:\{1...m\}$. Let us introduce **C**, a set of m (one for each hypothesis) integer counters initialized at 1. **C** is also indexed by ranks $i:\{1...m\}$. At each of the B data permutations, m checks are performed. Each check involves the evaluation of the first i hypotheses. Let $\check{T}_{(i)}$ be the maximal observed T-statistic across the i smaller statistics, that is,

$$\check{T}=\max(T_{(i)}),\ i:\{1...m\}. \qquad (5.0)$$

For example $\check{T}_{(1)}= T_{(1)}$; $\check{T}_{(2)}= \max(T_{(1)}, T_{(2)})$; $\check{T}_{(3)}= \max(T_{(1)}, T_{(2)},T_{(3)})$ etc.. $\check{T}$ can be conceived as a set of direct combining functions (CF) of the Tippett kind (Pesarin, 2001). Their null space is found permuting data and analogously finding at each data permutation the maximal statistic across the i smaller statistics. At each j^{th} permutation this would be

$$\check{T}_j^*=\max(T_{j(i)}^*),\ i:\{1...m\},\ j:\{1...B\}. \qquad (6.0)$$

For each data permutation the i^{th} counter is increased if $\check{T}_{j(i)}^* \geq \check{T}_{(i)}$. The check is performed each time for all m hypotheses. Once concluded the permutation process the p-value for each hypothesis is derived as $P_{(i)} = C_{(i)}/ B$. The procedure starts rejecting the m^{th} hypothesis and steps all way down to the first. At each step the hypothesis is rejected if the p-value is less than or equal to α. As soon as the p-value exceeds α the procedure

stops (enforced monotonicity to the p-values, see Westfall & Young, 1993), and all remaining hypotheses are accepted.

A desirable property of all procedures based on the maximum of test-statistic is that the power remains pretty constant varying the proportion of true null hypotheses, as it will be shown in the simulations. However when there is more than one false hypothesis the single-step procedure is overly conservative (Troendle, 1995) and the gain in power with step-down algorithms seems to be minimal (Blair and Karnisky, 1994; Holmes et *al.*, 1996). We now illustrate a generalization of the multiple comparisons permutation framework and a new procedure based on sums of (weighted) statistics instead that on maximal statistics.

The t- sum permutation test

The t-max combining function is here shown to be a particular case of a family of combining function which permutation space can be derived by resampling techniques without loss of generalization. In this sense the t-sum test can be considered as a generalization of the t-max test. For a t-sum test the (weighted) sum of the smaller i test-statistics is considered instead that of its maximum value. The relationship can be seen generalizing (5.0) and (6.0) as

$$\mathcal{T}=\Sigma_i(T_{(i)}W_{(i)}),\ i:\{1...m\}. \qquad (5.1)$$

and similarly for the permutation space

$$\check{T}_j^*=\Sigma_i(T_{j(i)}^*W_{(i)}),\ i:\{1...m\},\ j:\{1...B\} \qquad (6.1)$$

where W is a non-decreasing weighting function always applied to the sorted set of test-statistics. Setting $W_{(i)}=1$ if $k = m$, $W_{(i)}= 0$ otherwise, reduces (5.1) and (6.1) to (5.0) and (6.0) respectively. In this case the combining functions $\bar{T}$ and $\check{T}$ are equivalent and so is their null space. Setting $W_{(i)}=1$ for each i simplifies (5.1) and (6.1) to

$$\bar{T}=\Sigma_i(T_{(i)}),\ i:\{1...m\} \qquad (5.2)$$

and

$$\bar{T}_j^*=\Sigma_i(T_{j(i)}^*),\ i:\{1...m\},\ j:\{1...B\} \qquad (6.2)$$

leading to a permutational multiple test procedure using the (unweighted) sum of statistics (Pesarin, 2001, pag. 151). Computationally, the procedure we propose follows exactly the same steps as the t-max step-down procedure. Only the m combining functions and their null spaces are different. Using (5.2) and (6.2) instead of (5.0) and (6.0) respectively in the step-down all-at-once procedure described above leads to the (unweighted) t-sum procedure, which properties we next investigate by means of two sets of simulations. Table 1 reports the computational algorithm we employed.

**********Insert Table 1 approximately here**********

In the algorithm of table 1 the (corrected) threshold of significance is defined as the smallest observed statistic $T_{(q)}$ correspondent to $\mathbf{C}_q$ for which $\pi_q \leq \alpha$ holds. For example if it holds for q=5 then the threshold equals $T_{(5)}$. If $\pi_q \leq \alpha$ does not hold for any q we stop at stage 5). In this case the threshold of significance is undefined and the omnibus hypothesis $\boldsymbol{H^0}$ is accepted. The π_q are literally that: the probability under the null hypothesis of obtaining a (cumulative sum) test statistic for dimension q as large as the one observed. We see that expressions (5.1) and (6.1) are a more general form of an infinite family of combining functions based on the sum of test-statistics. It is known that the (equal weight) t-sum test on the m hypotheses controls the FWE in the weak sense (Arndt *et al.*, 1996; Blair and Karnisky, 1994; Karniski et *al.*, 1994). The t-sum test we propose using the step-down all-at-once procedure is an extension to any configuration of the m_0 and m_1. Furthermore, weights can be applied to refine the analysis, as it will be suggested in the discussion. We now turn to the empirical study of the properties of the equal weight version.

Simulation 1

Aim of the first set of simulations was to explore the properties of three multiple testing procedures under a wide range of conditions. We compared the t-max and t-sum (equal-weights) procedures in addition to simple univariate permutation tests (t-uni). t-max (step-down all-at-once) and t-sum as seen above make use of the same algorithm but different combining functions. The t-uni procedure is simply a series of partial test. For each procedure the power, FDR, SFE, FWE, and MSE were estimated by arithmetic means across up to 5000 repetitions of the same simulation. The error rate to be

controlled was set to 0.1 for all errors. Power was defined as the proportion of false hypotheses correctly rejected. With this definition the power is bounded between one and zero. Power equals zero means that none of the false hypotheses was rejected. Power equals one means that all false hypotheses were rejected. This quantity is independent from the m_0/ m_1 ratio meaning that it is a consistent estimator of the proportion of the false hypotheses correctly rejected *regardless* the actual proportion of true and false hypotheses. The simulations concerned a multidimensional one-sample student t framework where the central location of the distributions is compared to 0.0. Sample data under H^0 were generated so to be standard normal. The effect imposed under H^1 was fixed and equal to 1.0. We used gaussian data and fixed effects so to comply with the aforementioned MSE definition (2.0, 3.0). We varied the total number of variables (V), the sample size (N), the correlation structure of the data, and the proportion of activated variable (PAV). V used were 20, 100, 1000, and 10000. N were 8, 16, and 24, which cover the range of sample sizes more typically used in neuroimaging. Two level of data correlation were considered. In the first no correlation was enforced. In the other the correlation structure was homogeneous and so to be on the average 0.4. The proportion of activated variables (PAV) was defined respect to V. According to previous notation PAV = m_1/ m, while V = m. PAV used were 0, 0.05, 0.25, 0.5, 0.75, 1.0. For PAV = 0 the omnibus hypothesis (1.0) is true and all errors (FDR, SFE and FWE) are the same and refer to the control of the FWE in the weak sense. For correlated data we investigated only V=20 and V=100. All simulations used 5000 repetitions with the exception of simulations for V=1000 and 10000, for which we used 1000 repetitions. All simulations were carried out by a dedicated program written in Borland Delphi.

Results are presented in graphical form so to ease the comparisons of the simulations. In all charts the power, the FDR, the SFR, or the FWE are plotted on the y-axis against the level of PAV (x-axis). The t-max procedure is always represented by a dotted curve, t-sum by a solid curve, and t-uni by a dash-dot curve. To facilitate interpretation of results plots are clustered in a regular grid so that N varies along the horizontal dimension and V varies along the vertical dimension. Doing so, different figures show the same set of simulations varying the correlation structure. Figures 1 and 6, display the Power results with no correlation enforced and correlation equals 0.4 respectively. Figures 2 and 7, display the same results for the FWE (in the weak and strong sense), figures 3 and 8, for the FDR, and figures 4 and 9 for the SFE. Finally, figures 5 and 10 show the corresponding MSE estimations.

**********Insert figure 1 to 10 approximately here**********

Simulation 2

Aim of the second set of simulations was to compare the power and model fit of the t-sum and FDR procedure (Benjamini and Hochberg, 1995). Since the t-sum adopts a more lenient control of the type I error than the FDR, we were particularly interested in comparing the model fit of the two procedures. This set of simulation was produced by a computer program written independently from the previous, using the Mathworks Matlab suite. The simulations in this second set are a small subset of the first and use the

same parameters. Here we restricted ourselves to the case of no data correlation enforced. For both the power and the MSE it was produced a standard simulation with N=10, V=100, and fixed effect=1.0. In three further simulations one parameter at the time was varied; the sample size was raised to 20, the number of variables to 200, and the effect to 2.0. Results for power and MSE are reported in a graphical fashion as before. In all charts the t-sum procedure is always represented by a solid line, and the FDR procedure by a dotted line. Results of power and MSE are reported in figures 11 and 12, respectively.

**********Insert figure 11 and 12 approximately here**********

Results

Several conclusions can be drawn from the first set of simulations (figure 1 to 10):

1. As it is known both the t-max and the t-sum procedures control the FWE in the weak sense (see PAV=0 in figures 2 and 7) regardless the *n*/ *m* ratio and the correlation structure. The control is exercised at a level very close to the nominal one (see also Blair and Karnisky, 1994). On the other hand t-uni does not.

2. Whereas the t-max procedure controls the FWE also in the strong sense regardless the *n*/ *m* ratio and the correlation, and also controls the FDR and the SFE, the t-sum procedure controls the SFE but not the FDR nor the FWE in the strong sense.

However the actual proportion of false rejection of the t-sum procedure does not grow indefinitely. In the case of correlated data (figure 8) the FDR is actually always below the nominal level (0.1). See the section "FDR as noise estimator for a procedure controlling the SFE" for an interpretation of these results. On the other hand the t-uni procedure does not control any error considered. As a consequence it is not worth consideration as a multiple comparison procedure.

3. The t-sum procedure is substantially more sensitive then the t-max procedure in almost all situations considered. The gain in power increases with the proportion of activated voxels (PAV), as the sample size decreases, and as the number of hypotheses increases. The advantage of the t-sum procedure decreases as the correlation structure of the data increases.

4. In general, the power of the t-max procedure is approximately uniform with a slight positive trend across level of PAV. The power of the t-sum procedure, instead, increases as a function of PAV. For PAV=1.0, i.e., when all hypotheses are false, t-sum has always maximal power. While the power of the t-max procedure decreases as both the sample size decreases and the proportion of activated variables decreases, the power of the t-sum procedure remains almost constant across the levels of PAV.

5. The model fit as measured by the MSE is considerably better for the t-sum as compared to the t-max procedure. It is also good in absolute term, that is, it is always close to one. Interestingly, the t-sum MSE is also lower than the t-uni MSE in several

situations. Like for the power, the advantage of the t-sum procedure over the t-max procedure increases as the number of hypotheses increases and is reduced by data correlation.

Analogously several conclusions can be drawn from the second set of simulations (figure 11 and 10):

6. The t-sum procedure is more powerful than the FDR procedure in all case considered but when the effect is large.

7. This is not a consequence of the more lenient control over the type I error exercised by the t-sum procedure, in fact the results of the model fit are very similar, indicating that the t-sum procedure better maximizes the type I/ type II error trade-off.

Conclusions

According to our simulations (gaussian data with fixed effect) the power of the t-max procedure drops dramatically as the n/m ratio decreases. In the case of no correlation with sample size equals 8 and number of variables equals 10000 (figure 1) the power of the t-max procedure is close to zero. Following the trends in figure 6, and according to other simulations we performed and not reported here, the gain in power in the case of moderate correlation is not considerable. With 100000 data-points the

power of the t-max procedure would be extremely low even for moderately large sample sizes. While the t-sum is always more powerful under the complete falseness of H^0 and for some configurations of hypotheses wherein the ratio m_1/m_0 is high; decreasing the sample size and increasing the number of hypotheses leads fast to the situation where the t-sum is more powerful regardless the m_1/m_0 ratio. The t-sum is a more lenient procedure than the t-max. Indeed it controls only the FWE in the weak sense and the SFE. Under these circumstances power increase is not surprising. In fact power measures are affected by type II errors but not from type I errors, hence, typically they issues results somehow inversely related to the strictness of type I error control exercised. See for example the large power of the t-uni procedure. However the same is not true for the mean square error. The MSE is sensitive to both type I and type II errors and represents a measure of goodness of model fit. The very low MSE displayed by the t-sum procedure (figure 5 and 10) suggests that it optimizes the type I/ Type II error trade-off, a task truly at the core of the multiple testing problem. On the other hand, the large MSE of the t-max procedure confirms the known problem of this procedure, i.e., the fact that it is unduly conservative when several hypotheses are false. Indeed for the t-max procedure the larger the PAV the worse the MSE. However for sufficiently large n/m ratios (>0.2), the t-max procedure possesses all the desirable properties for a multiple comparison procedure, namely, high power (and uniform across PAV), low MSE, and strong control of the FWE. See for example the N=24 and V=100 case in figures 1 to 10.

Since the t-sum procedure is less stringent than the t-max procedure one may ask how it performs when compared to the FDR controlling procedure, which is also more lenient. The second set of simulations indicates that the t-sum may considerably

outperform the FDR procedure in power (figure 11) under several circumstances. The t-sum procedure fits our simple multivariate gaussian model better than the FDR (figure 12). It was shown in the previous set of simulations that the t-sum has better fit also when compared to the t-max procedure, and even when compared with a simple series of univariate tests.

Discussion

Experimentation in neuroimaging is usually expensive. For this reason the use of small sample sizes is not an exception. On the other hand the number of data-points (voxels) increases with the advance in neuroimaging technologies. The technological trend is upward; henceforth it is very important to device multiple testing procedures which power does not decrease as a function of the sample size/ number of hypothesis ratio. Under these and similar experimental circumstances the t-sum procedure offers more versatility than the t-max procedure. The t-sum procedure could be used in pilot and exploratory studies, when the sample size is usually very low, yet the optimization of the type I/ Type II error trade off is crucial. In these studies the researcher should be able to know if and where an effect exists, that is, sensitiveness in the presence of noise is sought. At the same time he/ she needs to be protected against false positive results, since more involving (and more expensive) studies may be planned as a consequence of preliminary results. Since the t-sum procedure has excellent model fit properties even in

noisy conditions, still controlling the FWE in the weak sense and the SFE, it is suitable for that purpose.

We have seen that the t-sum power increases as a function of PAV. Setting adequately (still a priori) the weighting function W may result into an optimal power procedure that adaptively account for the expected proportion of false hypotheses. Following the generalization (5.1) and (6.1), the t-sum and the t-max combining function can be conceived as being the two extreme points of a continuum. At one side W is a step function suppressing all but the maximal statistic (t-max). At the other, W weights equally all statistics (t-sum). One can conceive combining functions were the first $\hat{H}^0$ (estimated number of true hypotheses) with small test-statistic have zero weight and the remaining hypotheses with higher test-statistic have weight equal to 1. Existing methods for the estimation of number of true hypotheses (e.g., Turkheimer, Smith and Schmidt, 2001) could be used for this purpose. The weighting function does not need to be step-like. Logistic, exponential and other form of functions could be used as well. The selection of an optimal combining function from the (5.1) family for a given data-set is a new field of research and needs further investigation.

Figures and Tables

TABLE 1. ALGORITHM FOR THE T-SUM (EQUAL WEIGHT) PROCEDURE.

1) Set α and the number of random data permutations B

2) Initialize m integers counters to 1: $\mathbf{C}_i=1$; i:{1...m}

3) For the observed data compute the m combining functions (cumulative sums) on the observed data as $\underset{\sim}{T}=\Sigma_i(T_{(i)})$, i:{1...m}. (5.2)
For example: $\underset{\sim}{T}_{(1)}= T_{(1)}$; $\underset{\sim}{T}_{(2)}= T_{(1)} + T_{(2)}$; etc..

4) For every data permutation j compute $\underset{\sim}{T}_j^*=\Sigma_i(T_{j(i)}^*)$, (6.2)
and for each voxel i increase $\mathbf{C}_{(i)}$ if $\underset{\sim}{T}_{j(i)} > T_{(i)}$

5) Set $q=m$. This step corresponds to the Omnibus test as specified in (1.0).

6) Reject H^0_q if $\pi_q \leq \alpha$, where $\pi_q =(\mathbf{C}_q/B)$.

7) Set $q=q$-1 and return to step 6). The algorithm stops as soon as $\pi_q \leq \alpha$ does not hold.

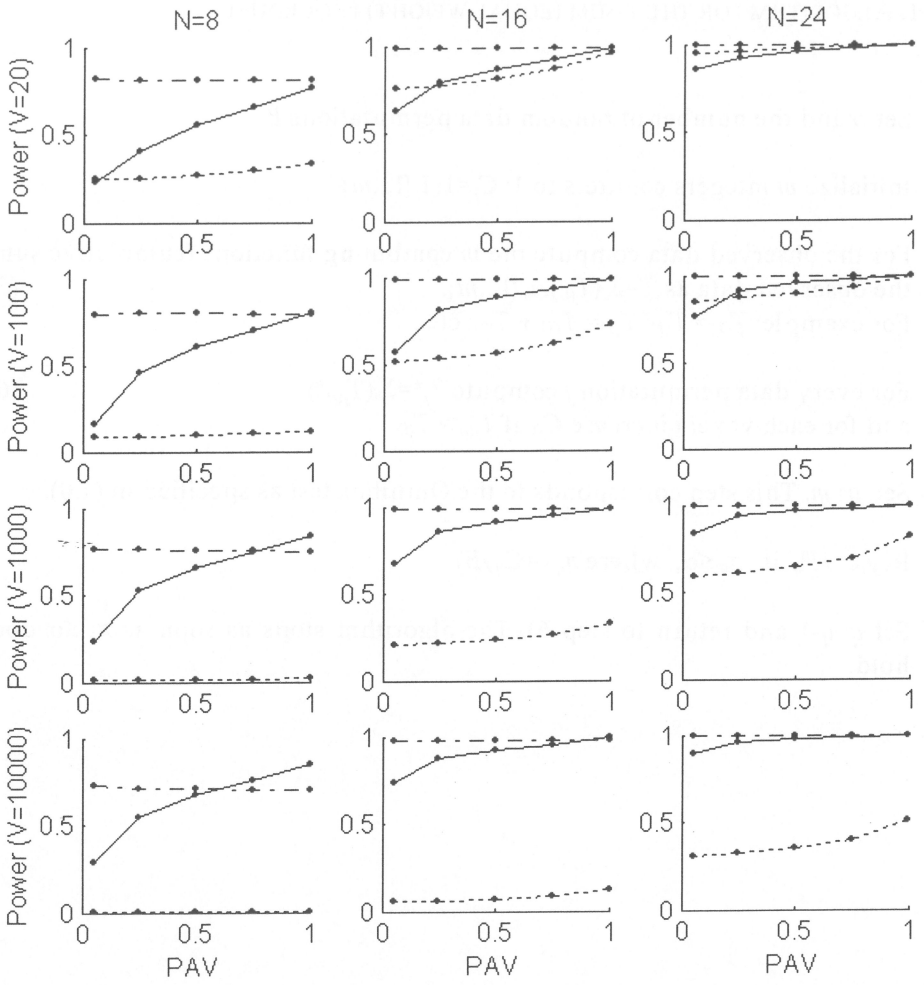

FIGURE 1. POWER OF T-MAX, T-SUM, AND T-UNI IN THE CASE OF NO CORRELATION. Power for t-max (dot line), t-sum (solid line), and t-uni (dot-dash line) is plotted over the proportion of active variables (PAV). From left to right it varies the sample sizes (8, 16, and 24). From top to bottom it varies the number of variables (20, 100, 1000, 10000).

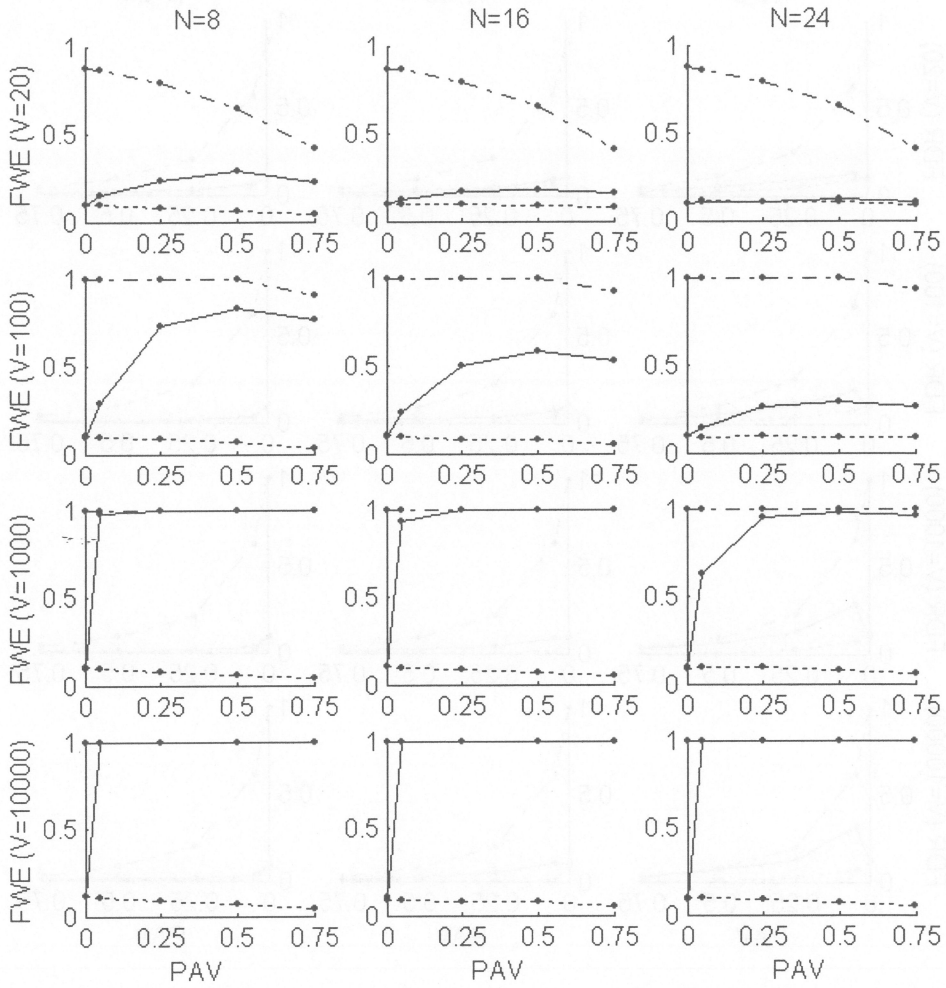

FIGURE 2. FAMILY-WISE ERROR OF T-MAX, T-SUM, AND T-UNI IN THE CASE OF NO CORRELATION. Errors for t-max (dot line), t-sum (solid line), and t-uni (dot-dash line) is plotted over the proportion of active variables (PAV). From left to right it varies the sample sizes (8, 16, and 24). From top to bottom it varies the number of variables (20, 100, 1000, 10000). For PAV=0 the error reported corresponds to the FWE in the weak sense (omnibus hypothesis), whereas for all other levels of PAV the error corresponds to the FWE in the strong sense.

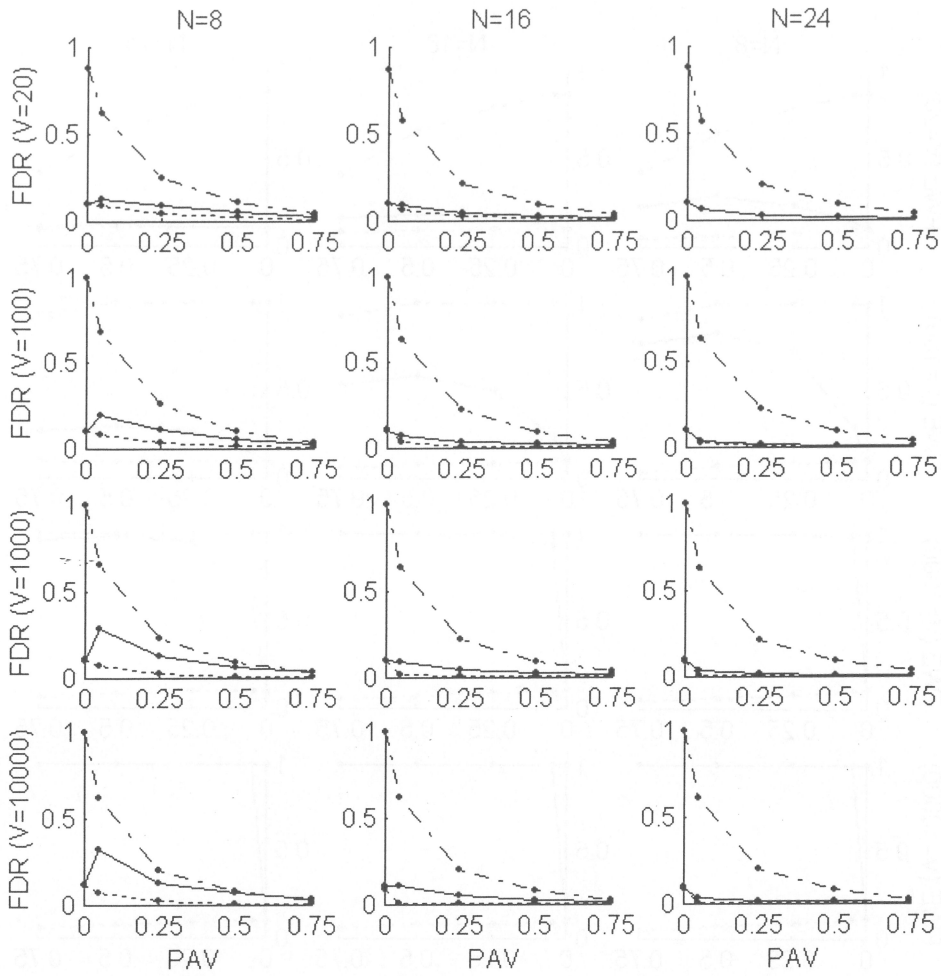

FIGURE 3. FALSE DISCOVERY RATE OF T-MAX, T-SUM, AND T-UNI IN THE CASE OF NO CORRELATION. Errors for t-max (dot line), t-sum (solid line), and t-uni (dot-dash line) is plotted over the proportion of active variables (PAV). From left to right it varies the sample sizes (8, 16, and 24). From top to bottom it varies the number of variables (20, 100, 1000, 10000). For PAV=0 the error reported equals the FWE error, hence it is the same as in Figure 2.

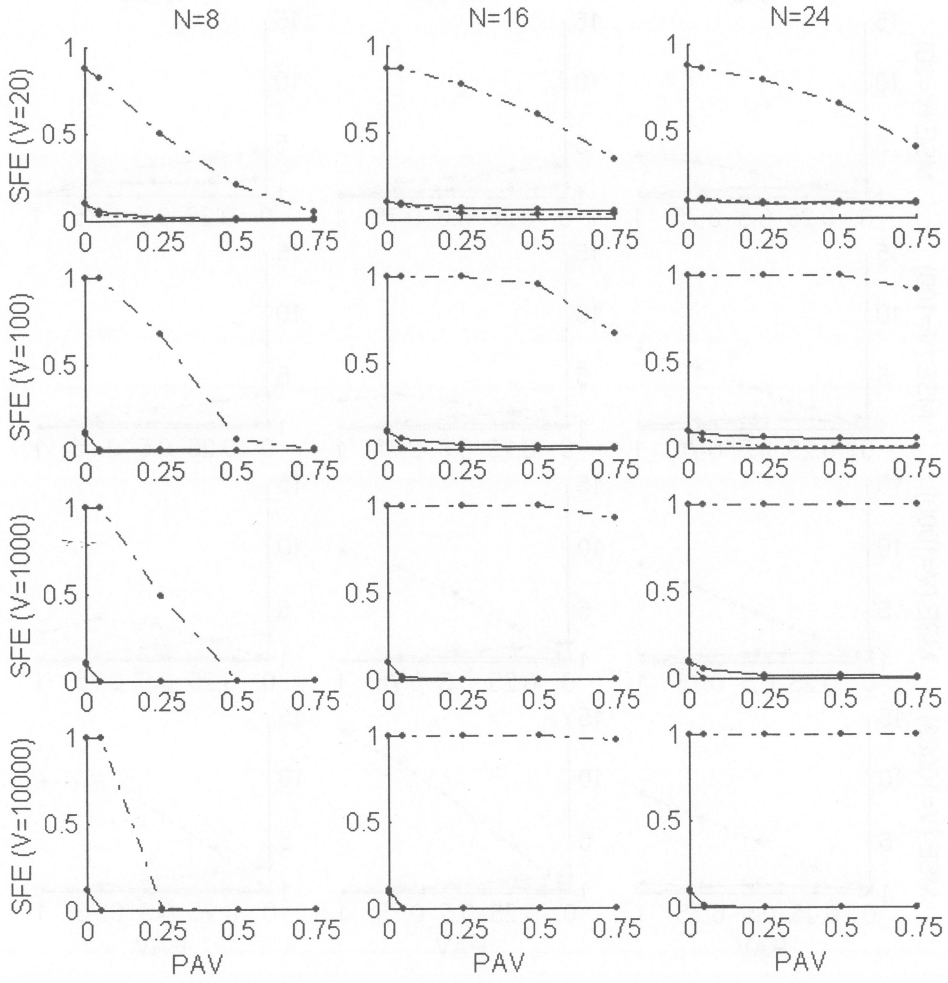

FIGURE 4. STOCHASTIC FAMILY ERROR OF T-MAX, T-SUM, AND T-UNI IN CASE OF NO CORRELATION. Errors for t-max (dot line), t-sum (solid line), and t-uni (dot-dash line) is plotted over the proportion of active variables (PAV). From left to right it varies the sample sizes (8, 16, and 24). From top to bottom it varies the number of variables (20, 100, 1000, 10000). For PAV=0 the error reported is equal to the FWE in the weak sense and to the FDR, hence it is the same as in Figure 2 and 3.

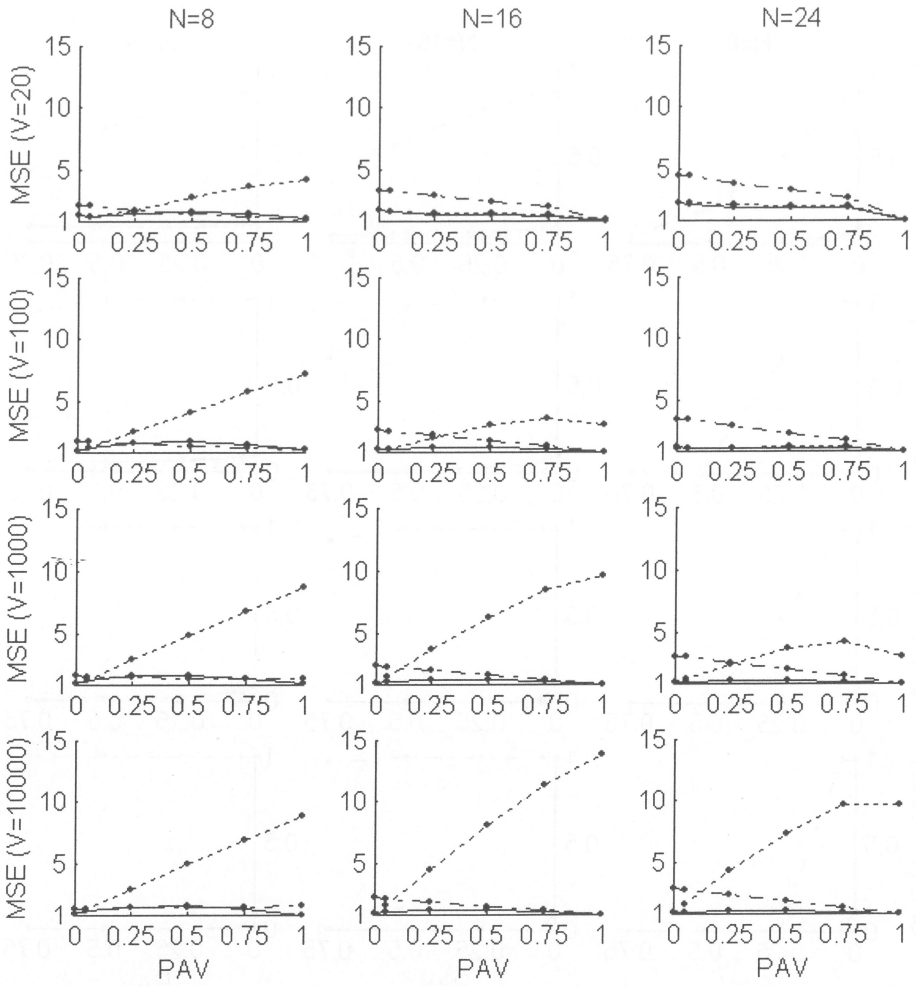

FIGURE 5. MEAN SQUARE ERROR OF T-MAX, T-SUM, AND T-UNI IN THE CASE OF NO CORRELATION. MSE for t-max (dot line), t-sum (solid line), and t-uni (dot-dash line) is plotted over the proportion of active variables (PAV). From left to right it varies the sample sizes (8, 16, and 24). From top to bottom it varies the number of variables (20, 100, 1000, 10000). MSE=1 corresponds to the theoretical perfect fit to the model. The greater the MSE, the poorer the fit to the model.

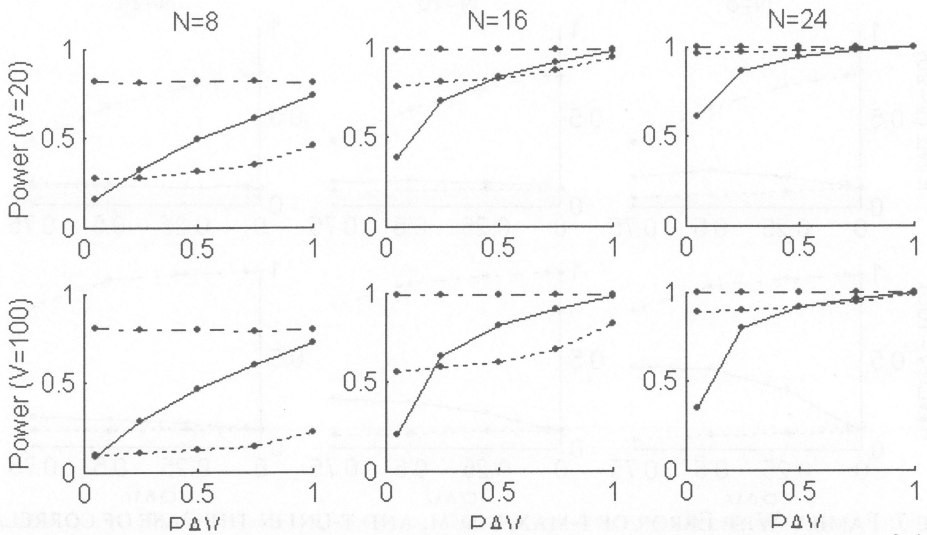

FIGURE 6. POWER OF T-MAX, T-SUM, AND T-UNI IN THE CASE OF CORRELATION =0.4. Power for t-max (dot line), t-sum (solid line), and t-uni (dot-dash line) is plotted over the proportion of active variables (PAV). From left to right it varies the sample sizes (8, 16, and 24). From top to bottom it varies the number of variables (20, 100, 1000, 10000).

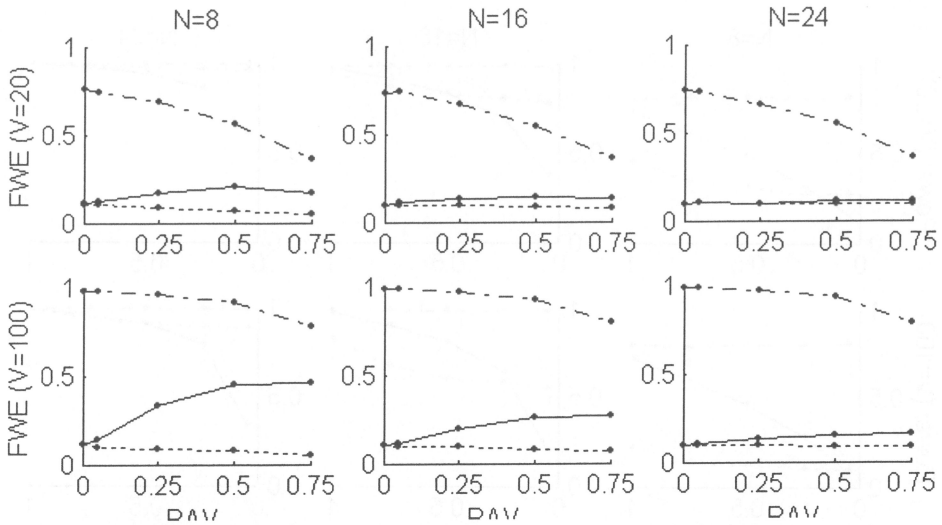

FIGURE 7. FAMILY-WISE ERROR OF T-MAX, T-SUM, AND T-UNI IN THE CASE OF CORRELATION =0.4. Errors for t-max (dot line), t-sum (solid line), and t-uni (dot-dash line) is plotted over the proportion of active variables (PAV). From left to right it varies the sample sizes (8, 16, and 24). From top to bottom it varies the number of variables (20, 100, 1000, 10000). For PAV=0 the error reported corresponds to the FWE in the weak sense (omnibus hypothesis), whereas for all other levels of PAV the error corresponds to the FWE in the strong sense.

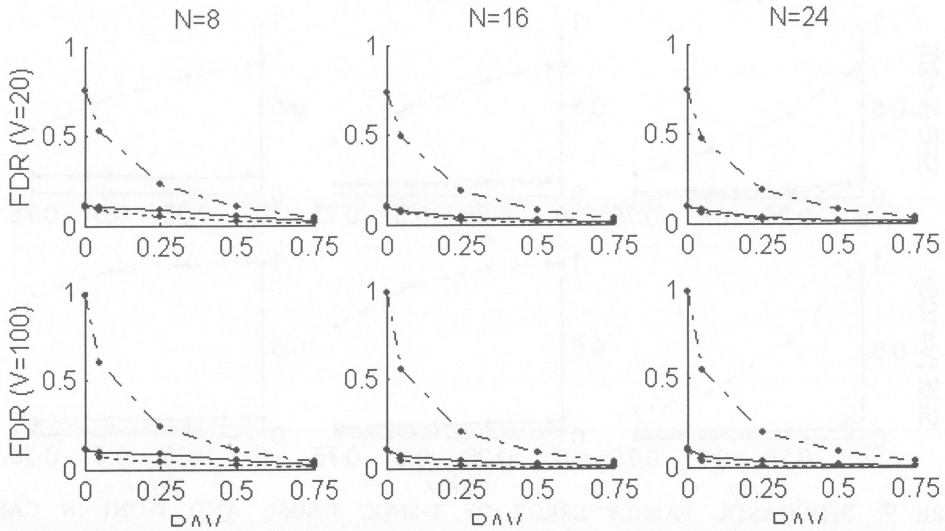

FIGURE 8. FALSE DISCOVERY RATE OF T-MAX, T-SUM, AND T-UNI IN THE CASE OF CORRELATION =0.4. Errors for t-max (dot line), t-sum (solid line), and t-uni (dot-dash line) is plotted over the proportion of active variables (PAV). From left to right it varies the sample sizes (8, 16, and 24). From top to bottom it varies the number of variables (20, 100, 1000, 10000). For PAV=0 the error reported equals the FWE error, hence it is the same as in Figure 2.

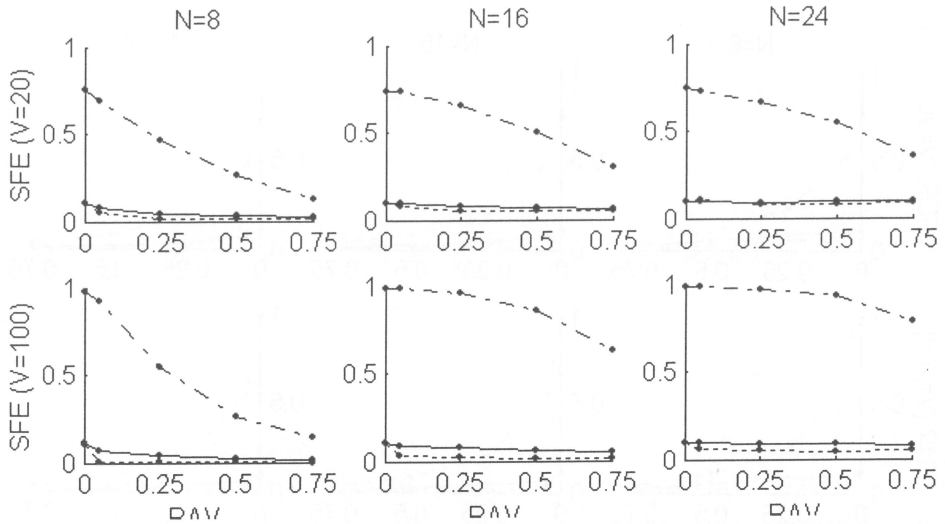

FIGURE 9. STOCHASTIC FAMILY ERROR OF T-MAX, T-SUM, AND T-UNI IN CASE OF CORRELATION =0.4. Errors for t-max (dot line), t-sum (solid line), and t-uni (dot-dash line) is plotted over the proportion of active variables (PAV). From left to right it varies the sample sizes (8, 16, and 24). From top to bottom it varies the number of variables (20, 100, 1000, 10000). For PAV=0 the error reported is equal to the FWE in the weak sense and to the FDR, hence it is the same as in Figure 2 and 3.

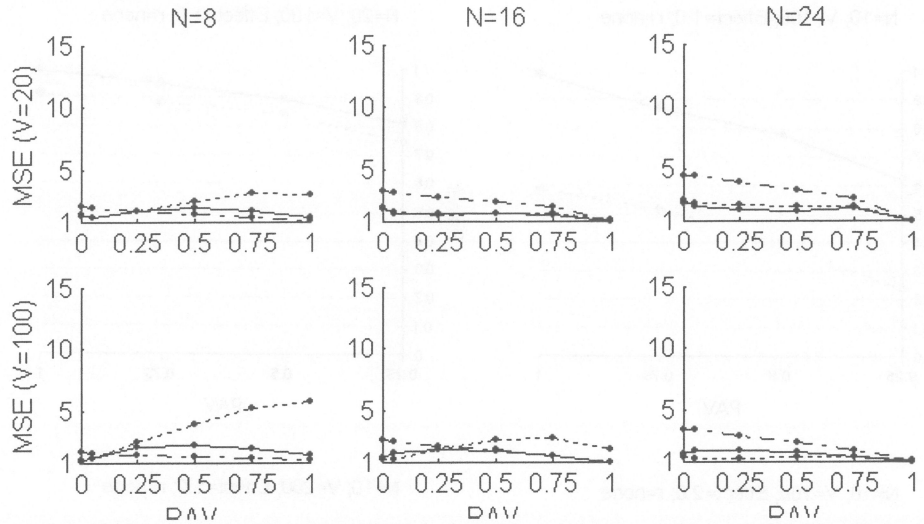

FIGURE 10. MEAN SQUARE ERROR OF T-MAX, T-SUM, AND T-UNI IN THE CASE OF CORRELATION =0.4. MSE for t-max (dot line), t-sum (solid line), and t-uni (dot-dash line) is plotted over the proportion of active variables (PAV). From left to right it varies the sample sizes (8, 16, and 24). From top to bottom it varies the number of variables (20, 100, 1000, 10000). MSE=1 corresponds to the theoretical perfect fit to the model. The greater the MSE, the poorer the fit to the model.

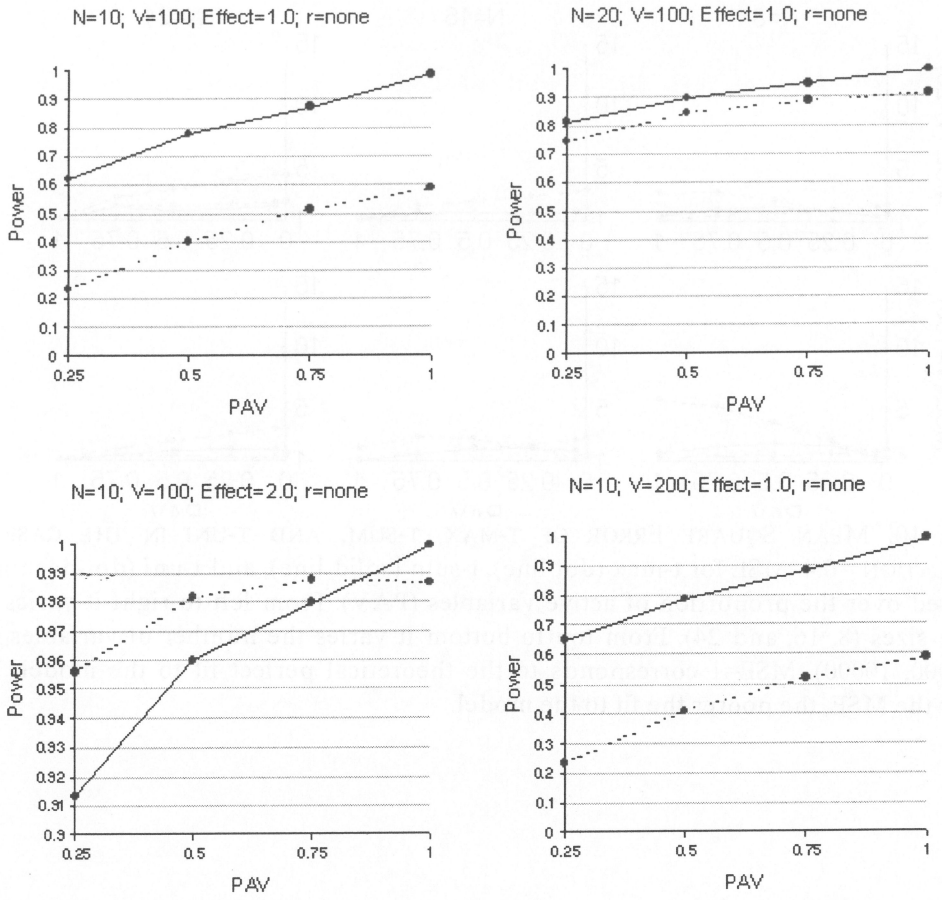

FIGURE 11. POWER OF T-SUM AND FDR. Power for t-sum (solid line), and FDR (dot line) is plotted over the proportion of active variables (PAV). The chart at the left-upper corner represents a standard simulation with sample size (N)=10, 100 variables (V), and effect =1.0. In the other charts one parameter at the time has been varied. All data were simulated without enforcing correlation.

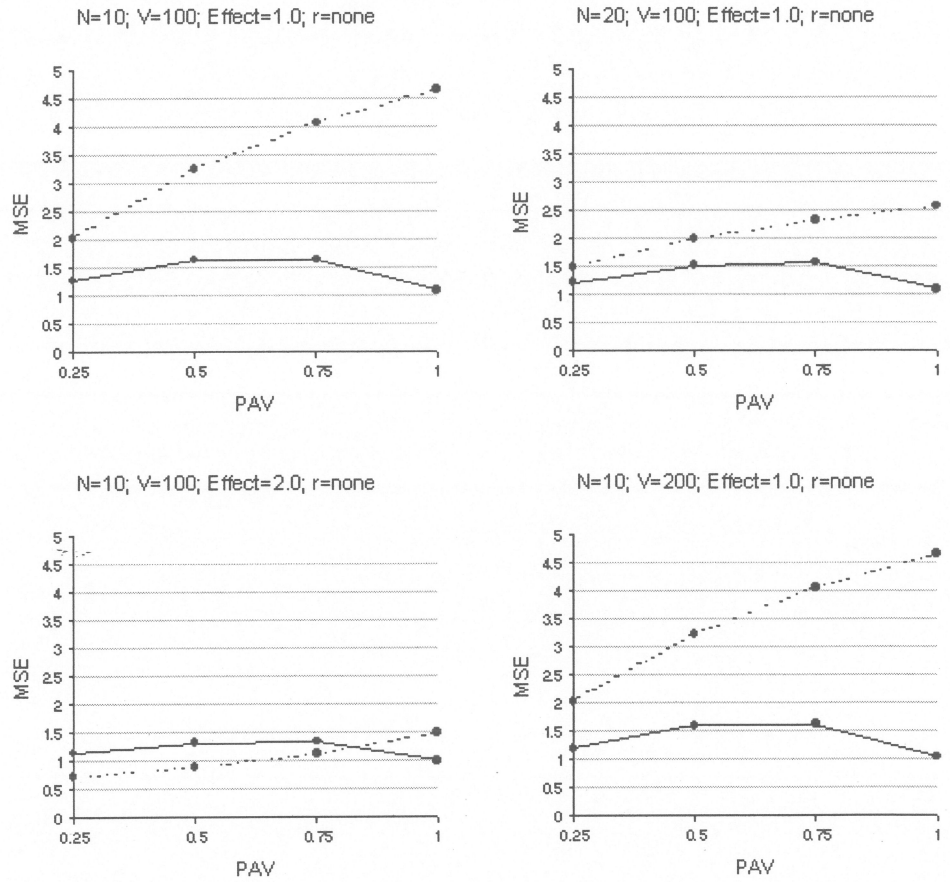

FIGURE 12. MSE OF T-SUM AND FDR. Mean square error of t-sum (solid line), and FDR (dot line) is plotted over the proportion of active variables (PAV). The chart at the left-upper corner represents a standard simulation with sample size (N)=10, 100 variables (V), and effect =1.0. In the other charts one parameter at the time has been varied. All data were simulated without enforcing corrrelation. MSE=1 corresponds to the theoretical perfect fit to the model. The greater the deviation from 1.0, the poorer the fit to the model.

References

Abramovich, F., Benjamini, Y., Donoho, D., Johnstone, I., 2000. Adapting to Unknown Sparsity by controlling the False Discovery Rate, Stanford Statistics Department, Technical Report # 2000-19 (available at http:/ / www.math.tau.ac.il/ ~felix/ Papers.html).

Arndt, S., Cizadlo, T., Andreasen, N. C., Heckel, D., Gold, S., O'Leary, D. S., 1996. Tests for Comparing Images Based on Randomization and Permutation Methods. Journal of Cerebral Blood Flow and Metabolism 16, 1271-1279.

Bellman, R., 1961. Adaptive Control Processes: A Guided Tour. Princeton University Press.

Belmonte, M., Yurgelun-Todd, D., 2001. Permutation testing Made Practical for Functional Magnetic Resonance image Analysis, IEEE Transactions on Medical Imaging 20 (3), 243-248.

Benjamini, Y., Hochberg, Y., 1995. Controlling the False discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society B 57 (1), 289-300.

Blair, R. C., Karnisky, W., 1994. Distribution-free Statistical Analyses of Surface and Volumetric Maps. In: R. W. Thatcher, M. Hallett, E. R. John, M. Huerta (Eds.), Functional Neuroimaging: Technical Foundations. Academic Press, San Diego.

Blair, R. C., Troendle, J. F., Beck, R. W., 1996. Control of Familywise Errors in Multiple Assessments via Stepwise Permutation Tests, Statistics in Medicine 15, 1107-1121.

Bullmore, E., Brammer, M., Williams, S. C. R., Rabe-Hesketh, S., Janot, N., David, A., Mellers, J., Howard, R., Sham, P., 1996. Statistical methods of estimation and inference for functional MR image analysis, Magnetic Resonance in Medicine 35 (2), 261-277.

Edgington, E. S., 1995. Randomization tests. 3rd ed. Marcel Dekker, New York.

Feinstein, A. R., 1993. Permutation tests and "statistical significance". MD Computing 10 (1), 28-41. (reprint of a book chapter published by C.V. Mosby in 1977).

Fisher, R. A., 1935. Design of Experiments. ed. Oliver and Boyd, Edinburgh.

Genovese, C. R., Lazar, N. A., Nichols, T., 2002. Thresholding of statistical maps in functional neuroimaging using the false discovery rate. Neuroimage 15 (4), 870-878.

Hochberg, Y., Tamhane, A., 1987. Multiple Comparisons Procedures, John Wiley & Sons, New York.

Holmes, A. P., Blair, R. C., Watson, J. D. G., Ford, I., 1996. Nonparametric Analysis of Statistic Images from Functional Mapping Experiments. Journal of Cerebral Blood Flow and Metabolism. 16 (1), 7-22.

Karniski, W., Blair, R. C., Snider, A. D., 1994. An Exact Statistical Method for Comparing Topographic Maps, with any Number of Subjects and Electrodes. Brain Topography 6 (3), 203-210.

Lunneborg, C. E., 2000. Data Analysis by Resampling: Concepts and Applications, Duxbury, Pacific Grove, California.

Manly, B. F. J., 1997. Randomization, Bootstrap and Monte Carlo Methods in Biology, 2nd ed. Chapman & Hall, London.

Nichols, T. E., Holmes, A. P., 2001. Nonparametric Permutation Tests for Functional Neuroimaging: A Primer with Examples. Human Brain Mapping 15, 1-25.

Pitman, E. J. G., 1937a. Significance tests which may be applied to samples from any population. Journal of the Royal Statistical Society B. 4, 119-130.

Pitman, E. J. G., 1937b. Significance tests which may be applied to samples from any population. II. The correlation Coefficient. Journal of the Royal Statistical Society B. 4, 225-232.

Pitman, E. J. G., 1938. Significance tests which may be applied to samples from any population. III. The analysis of Variance test. Biometrika 29, 322-335.

Pesarin, F., 2001. Multivariate Permutation tests. John Wiley & Sons, New York.

Petersson, K. M., Nichols, T. E., Poline, J-B, Holmes, A. P., 1999. Statistical limitations in functional neuroimaging II. Signal detection and statistical inference. Philosophical Transaction of the Royal Society of London 354, 1261-1281.

Sabatti, C., Karsten, S. L., Geschwind, D. H., 2002. Thresholding rules for recovering a sparse signal from microarray experiments. Math. Biosci. 176, 17–34.

Swets, J. A., 1988. Measuring the Accuracy of Diagnostic Systems. Science 240, 1285-1293.

Swets, J. A., Pickett, R. M., 1982. Evaluation of Diagnostic Systems. Methods from Signal Detection Theory. Academic Press, New York.

Troendle, J. F., 1995. A Stepwise resampling Method of Multiple Hypothesis testing, Journal of the American Statistical Association 90 (429), 370-378.

Troendle, J. F., 1996. A Permutation Step-up Method of Testing Multiple outcomes, Biometrics 952, 846-859.

Turkheimer, F. E, Smith, C. B., Schmidt, K., 2001. Estimation of the number of "true" null hypotheses in multivariate analysis of neuroimaging data, Neuroimage 13(5), 920-30.

Westfall P. H., Young S. S., 1993. Resampling-Based Multiple Testing. Examples and Methods for p-Values Adjustment. John Wiley & Sons, New York.